DOCUMENT RESUME

ED 289 913 TM 870 749

TITLE	New Jersey Statewide Testing System: High School Proficiency Test, 1985-86. Technical Report.
INSTITUTION	New Jersey State Dept. of Education, Trenton.
PUB DATE	86
NOTE	49p.; Developed by National Computer Systems and Beck Evaluation & Testing Associates, Inc. For related documents, see TM 870 747-752.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Equated Scores; Graduation Requirements; High Schools; Item Analysis; Mathematics Tests; *Minimum Competency Testing; Pretesting; Reading Tests; *Scores; *State Programs; Statewide Planning; *Test Construction; Testing Programs; Test Items; *Writing Evaluation
IDENTIFIERS	New Jersey; *New Jersey High School Proficiency Test

ABSTRACT

This Technical Report provides descriptions and summary data that assist measurement specialists in assessing the procedures used in developing New Jersey's High School Proficiency Test (HSPT), the technical qualities of the tests, and the statewide results obtained from its use. The data summarized in this report were collected during the various phases of development of the third annual form of the New Jersey High School Proficiency Test (HSPT) during the 1985-86 school year. The HSPT is administered annually to all public school students in grade nine and constitutes a requirement for high school graduation. The three major data collection phases were Item Development and Review, Field Testing, and HSPT results. The first three sections of this report are organized according to these three phases. The fourth section describes the equating process used to link the HSPT back to the first administration of the test in 1984. The final section discusses the development and administration of the essay test, as well as the procedures for deriving the total writing test score. (KSA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. Vespucci

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

NEW JERSEY STATEWIDE TESTING SYSTEM

HIGH SCHOOL PROFICIENCY TEST

1985 - 1986

TECHNICAL REPORT

Saul Cooperman
Commissioner of Education

Division of General Academic Education
Joel Bloom
Assistant Commissioner

Developed under the Direction of the
Bureau of Cognitive Skills

New Jersey Department of Education
225 West State Street
CN 500
Trenton, New Jersey 08625

Developed by

National Computer Systems

and

Beck Evaluation & Testing Associates, Inc.

STATE BOARD OF EDUCATION

JOHN T. KLAGHOLZ.....Seaside Park
President

JAMES JONES.....Teaneck
Vice-President

S. DAVID BRANDT.....Cherry Hill

MAUD DAHME.....Annandale

BETTY A. DEAN.....East Orange

MATEO F. DeCARDENAS.....Union City

ANNE S. DILLMAN.....Perth Amboy

ALICE A. HOLZAPFEL.....Elizabeth

NANCY SCHAENEN.....Madison

JAMES M. SEABROOK, SR. .....Bridgeton

ROBERT A. WOODRUFF.....Elmer

DEBORAH P. WOLFE.....Cranford
Representing the State Board of Higher Education

Saul Cooperman, Commissioner of Education
Secretary

T. Edward Hollander, Chancellor of Higher Education
Ex-Officio

Copyright © 1986 New Jersey State Department of Education

CONTENTS

Introduction i

Section I. Item Development and Review 1

Table I-1. Number of Objectives and Items by Test and Cluster Reviewed by Committees, Field Tested, and Included on the Final Versions of the 1986 HSPT 3

Table I-2. Summary of Revisions Made to Field-Tested Items During the Item Review Meetings--By Content Area According to the Extent and Type of Change Made . 4

Table I-3. Readability Information on the Passages Included on the HSPT Reading Test. 5

Section II. Field Testing. 6

Table II-1. Field Test Sample Sizes by Test, Form, Ethnic Group and Sex 9

Table II-2. HSPT Raw Scores for Field Test Samples According to Field Test Form Taken. 10

Table II-3. Summary Field Test Data by Content Area--Item Difficulty and Discrimination. 11

Table II-4. Summary of Field Test Data Analysis of Items with Selected Statistical Characteristics 12

Table II-5. Summary of Item Revisions Following Field Testing--By Test and Type of Change 13

Table II-6. Percentage of Omits in Field Test Items by Test and Ethnic Group 14

Section III. HSPT Results. 15

Table III-1. Total Population of Students Tested Statewide by District Factor Grouping. 18

Table III-2. Summary of Number of Students Tested or Excluded by Test Area on the HSPT. 19

Table III-3. Summary Data for the Total Statewide Population and the Subsample Used for Subsequent Statistical Analyses. 20

Table III-4. Raw Score Frequency Distributions and Summary Data by Test Cluster--Percent of Students Receiving Each Possible HSPT Raw Score 21

Table III-5a. Frequency Distributions of Item Difficulty Values and Point-Biserial Discrimination Indices by Cluster HSPT Reading 22

Table III-5b. Frequency Distributions of Item Difficulty Values and Point-Biserial Discrimination Indices by Cluster HSPT Writing 23

Table III-5c. Frequency Distributions of Item Difficulty Values and Point-Biserial Discrimination Indices by Cluster HSPT Mathematics 24

Table III-6. Raw Score Frequency Distributions and Summary Data by Test Percent of Students Receiving Each Possible HSPT Raw Score. 25

Table III-7. HSPT Summary Data and Reliability Estimates by Cluster and Test 26

Table III-8. Summary of the Consistency of HSPT Items with the Assumptions of the Rasch Model. 26

Table III-9. Summary Data for Number of Items Right, Wrong, Omitted or Not Reached by Part for the HSPT . . 27

Table III-10. Intercorrelation Matrix by Cluster and Test . . 28

Section IV. Equating 29

Table IV-1. 1986 Raw Score to HSPT Score Conversion Tables. 33

Section V. Essay and Total Writing Test 34

Table V-1. Distribution of HSPT Essay Scores 38

Table V-2. Distribution of HSPT Total Writing Test Score . 40

INTRODUCTION

This Technical Report summarizes data collected during the various phases of development of the third annual form New Jersey High School Proficiency Test (HSPT) during the 1985-86 school year. The three major data collection phases were Item Development and Review, Field Testing, and HSPT Testing. The first three sections of the Technical Report are organized according to these three phases. The fourth section of this report describes the equating process used to link the 1986 HSPT back to the first administration of the test in 1984. The final section discusses the development and administration of the essay test and the procedures for deriving the total writing test score.

The Technical Report is designed to provide descriptions and summary data that assist measurement specialists in assessing the procedures used in developing the HSPT, the technical qualities of the tests, and the statewide results obtained from its use. However, readers with a general knowledge of basic statistics and educational measurement should have little difficulty understanding most of the text and summaries provided here.

Readers desiring additional information concerning the statewide results based on the HSPT should obtain a copy of State Report: 1985-86 Analysis and Interpretation of Statewide Performance by contacting the Bureau of Cognitive Skills at 609-292-5180.

SECTION 1

Item Development and Review

The 1985-86 HSPT item development activities began with a review of the document entitled Skill Arrays for the New Statewide Testing System for each of the three tested content areas. Following this review by the contractor, discussions were held between the contractor and the Bureau of Cognitive Skills staff to review the HSPT skills, clarify the mode of assessment of various skills, and agree upon which skills should be assessed. On the basis of these discussions, it was decided to write items for 11 of the 12 Writing objectives (sufficient numbers of essay topics had been field tested the previous year to provide topics for this year's test), for 32 of the 35 Reading objectives (the remaining objectives cannot be adequately assessed with the HSPT format), and for all of the 45 Mathematics objectives.

Table I-1 lists the number of HSPT skills by test area and cluster, as well as the numbers of items written for review, approved for field testing, and included on the final HSPT tests. In general, 9 or 10 items were written for each Writing objective, from which 7 were field tested, and 6 included on the final HSPT. In Mathematics, varying numbers of items per objective were selected by the Mathematics committee for use in the final HSPT. These numbers were used to guide the item writing efforts. In all cases, more than a 50% excess number of items were written than were intended for use in the HSPT. In Reading, because of the dependence of most of the items on reading passages, 6 to 11 items per objective were prepared for review; these items were based on 25 passages. At least 4 items per objective were field tested; the final HSPT Reading test contains a constant 3 items per objective. The excess items in each area were developed to expedite the committee review process, to permit wider discretion for the committees in item selection and, where test specifications so permitted, to illustrate different ways to assess a given objective.

A total of 596 items were prepared for inspection by the Bureau of Cognitive Skills staff and the test development and minority review committees. These committees met for three days to discuss the items, make revisions they thought would improve the items, and select a subset of the items for field testing. Table I-2 provides a summary of the revisions made during this process to all items selected for the Field Test. Approximately one-third of the submitted items were revised; the most typical changes were to revise one of the answer options or to make a minor change in the wording of the item stem (question). Only 6% of the selected items required a major revision, and only 6 additional items were requested by the review committees to replace items originally submitted.

All reading passages included in the Reading test were developed to have estimated readability levels between Grades 8 and 10, as determined by the Dale-Chall readability formula. Since this formula provides estimates only

for pairs of grades at this level, a Dale-Chall estimate of Grades 7-8 was considered as being within this range. The reading difficulty of each passage was further examined subjectively for appropriateness by the Reading test development committee members. A summary of the readability formula data for the 16 passages included in the HSPT is provided in Table I-3. These results are based on the passages as presented in their final form on the HSPT. All passages fell within the state-specified grade range using the Dale-Chall criteria. It is important to note that readability formulas, while of some relevance, are widely criticized by reading experts. The "difficulty level" of a reading passage is an elusive construct only partially captured by even the best of readability formulas. For that reason, while the formula data were considered, the test development committee's judgment of the passage difficulty was considered as being at least equally important in making this determination. Even more important in this regard were the results of student performance by item and passage during the field testing.

In addition to the inspection of all the items by the three subject-matter test development committees, a minority advisory committee reviewed all items selected for field testing. All suggested revisions to these items suggested by this committee were incorporated into the field test materials.

Following the test development and minority committee meetings, selected items were organized into sections of appropriate length by the contractor, graphics created, and items sequenced for field testing. The items to be field tested were arranged into various forms to permit field testing within approximately 25-35 minutes. That is, rather than administering all 85 Writing items, for example, to the same students, these items were split into 4 separate forms of 21-22 items each to permit testing within a single class period. This procedure reduces the intrusion of the Field Test in the regular school program, while broadening the number of districts providing data to the Field Test process. Columns 3 and 4 of Table I-1 summarize the Field Test content by test and cluster relative to that of the final HSPT booklet. In all areas, more items were field tested than were used or intended for use in the final test. This was done to permit some degree of freedom for the development committees following field testing to select those items that were functioning as expected and intended, without requiring extensive rewriting at that stage. This is especially critical in the passage-based Reading test section. The success of this strategy is evident in the next section of this report, where data are presented on the small number of items requiring post-Field Test revision. Revision of test items prior to large-scale statewide use should be minimized to the extent possible to reduce the possibility of including weak or faulty items.

Table I-1

NUMBER OF OBJECTIVES AND ITEMS BY TEST AND CLUSTER REVIEWED BY COMMITTEES, FIELD TESTED, AND INCLUDED ON THE FINAL VERSIONS OF THE 1986 HSPT

AREA/Cluster	Number of Objectives	Number of Test Items		
		Reviewed	Field Tested	Final HSPT
READING	32	284	132	96
Literal Comprehension	11	107	45	33
Inferential Comprehension	13	120	55	39
Study Skills	8	57	32	24
WRITING	12	104	85	67
Sentence Structure	5	50	42	30
Organization of Ideas	4	36	28	24
Editing	2	18	15	12
Essay	1	--*	--*	1
MATHEMATICS	45	208	129	107
Fractions	5	28	15	15
Decimals	3	16	9	9
Percents	3	14	9	6
Concepts	9**	29**	19**	16
Measurement & Geometry	11+	40	28	20
Pre-Algebra	4	23	13	12
Problem Solving	10	58	36	29

\+ Items for a 12th objective were reviewed and field tested but then deleted.

* Essay topics were field tested in 1984.

** Two objectives (four items) were reviewed and field tested at separate times from the other items.

Table I-2

SUMMARY OF REVISIONS MADE TO FIELD-TESTED ITEMS DURING THE ITEM REVIEW MEETINGS -- BY CONTENT AREA ACCORDING TO THE EXTENT AND TYPE OF CHANGE MADE

Type/Extent of Revision	READING ITEMS		WRITING ITEMS		MATHEMATICS ITEMS		TOTAL ITEMS	
	Number	Percent	Number	Percent	Number	Percent	Number	Percent
No Change	105	80%	36	46%	85	66%	226	66%
1 option changed	11	8%	6	8%	12	9%	29	9%
2 options changed	3	2%	1	1%	1	1%	5	1%
3 options changed	0		0		0		0	
4 options changed	0		0		2	2%	2	1%
Stem - Minor change	6	5%	14	18%	10	8%	30	9%
Fewer than 5 words changed	1	1%	6	8%	1	1%	8	2%
One sentence changed	0		2	3%	0		2	1%
Two minor changes	3	2%	5	6%	3	2%	11	3%
Major change or rewrite	3	2%	9	11%	9	7%	21	6%
New items written/requested	0		0		6*	5%	6	2%
Total Number of Items Approved	132		79		129		340	

READING PASSAGES

Approved For Field Testing	17
Rejected	8

Of Approved Field Test Passages:

	No.	%
No Changes	7	41%
Fewer than 5 words changed	6	35%
One sentence changed/dropped	2	12%
Changes in more than 2 sentences	2	12%

*Four of these items were for a newly created objective, later dropped.

Table I-3

READABILITY INFORMATION ON THE PASSAGES
INCLUDED ON THE HSPT READING TEST

Passage	Topic	Number of Paragraphs	Number of Sentences	Average Sentence Length	% Words Off-List	Dale-Chall Readability Score	Dale-Chall Readability Grade
A.	Icebergs	5	18	15.72	16.96	7.09	Gr. 9-10
B.	Walt Whitman	5	18	19.11	18.90	7.57	Gr. 9-10
C.	Louisville Slugger	5	19	21.21	11.66	6.53	Gr. 7-8
D.	New Jersey	3	13	13.92	15.47	6.77	Gr. 7-8
E.	Dr. Mary Walker	5	19	21.63	17.76	7.53	Gr. 9-10
F.	Cape May	5	20	21.25	20.47	7.92	Gr. 9-10
G.	John Stevens	5	19	21.16	12.19	6.61	Gr. 7-8
H.	Bicarbonate of Soda	4	15	18.27	14.23	6.79	Gr. 7-8
I.	Battleship New Jersey	4*	16	22.88	16.39	7.36	Gr. 9-10
J.	Swordfish	5	27	14.41	10.80	6.06	Gr. 7-8
K.	Sara Vaughan	3	12	21.17	20.47	7.92	Gr. 9-10
L.	Harold's Job Search	5	25	12.40	13.23	6.34	Gr. 7-8
M.	Atlantic City Piers	4	21	18.90	16.88	7.24	Gr. 9-10
N.	Johnson Atelier	5	21	17.19	20.22	7.68	Gr. 9-10
O.	Elizabeth Haddon	4	15	17.60	15.91	7.02	Gr. 9-10
P.	Smoked Eel	6	23	16.56	16.54	7.07	Gr. 9-10

Summary of approximate readability grade levels:

Grade 7 - 2 passages
Grade 8 - 4 passages
Grade 9 - 5 passages
Grade 10 - 5 passages

*Plus a table.

SECTION II

Field Testing

All Field Tests were administered by State Department of Education personnel between April 17, 1986 - May 2, 1986. In most cases, testing occurred within the first week after the 1985 HSPT testing had been completed. Table II-1 summarizes the number of students included in the Field Test. Since Field Test data were analyzed separately by sex and ethnic group, as well as for the complete sample, the table shows the numbers of cases in each subgroup. The Field Test samples were selected on the basis of demographic factors to provide a close approximation of the statewide population and performance. Later analyses of the Field Test results vis à vis the Grade 9 results on the HSPT indicated that the Field Test sample provided a very close approximation to the total statewide population. Statewide data are provided in the 1985-1986 State Report, available from the Bureau of Cognitive Skills.

Items to be field tested were divided into several separate forms so the amount of testing time required per student would be minimized. Each Field Test form (6 in Reading, 5 in Mathematics, and 4 in Writing) was designed so it could be administered within approximately 30-40 minutes. Table II-1 shows the number of students who were tested with each Field Test form. After field testing, all items were reviewed for use on the HSPT independent of their Field Test form. Because of this, it was important to ensure that the Field Test samples taking each content area were of comparable ability. Table II-2 demonstrates the equivalence of these samples. Data summarized in this table are the 1985 HSPT raw scores for the Field Test samples. HSPT scores are for the comparable content area tests. It can be seen from these data that the several Field Test samples were closely comparable in ability, thus supporting the review of the Field Test items independent of the specific form in which they were field tested.

Results of the Field Test are summarized in Table II-3. The top portion of this table provides a frequency distribution of the item difficulty p-values for all items by test, as well as corresponding summary data. It can be seen from these results that some items from each of the tests were difficult for many students. This was especially true of the Mathematics test, for which the typical item was answered correctly by about two-thirds of the students tested. All three tests showed a wide range of item difficulty.

The bottom portion of Table II-3 summarizes the item discrimination values by test for the Field Test sample. These discrimination indices are the point-biserial correlation of the item score (correct or incorrect) with the test raw score. As a general rule of thumb, point-biserial indices for soundly developed competency tests tend to be in the .30-.50 range. Values below .30 often indicate items that are either extremely easy or difficult or that require further inspection and, perhaps, revision. The large majority of Field Test items (91%) yielded discrimination indices above the .30 level. It should be noted that low discrimination values often are found for items that the large majority of test takers answer correctly. If such items are judged to match the intended objective, they are retained in the test for content validity reasons.

All Field Test items were analyzed statistically in several ways. These analyses were conducted to assist the test development and minority review committees in their selection of items to be included on the HSPT. One set of analyses was conducted to determine the consistency of the items with the assumptions of the "Rasch model." This mathematical model is being used to equate successive versions of the HSPT. Since the equating of results from year to year is such a critical component of any state testing program, it is important that the test being built contain items that are consistent with the model. Two analyses were made to examine the match of the items with the assumptions of the model. Results are shown in Table II-4. First, and most importantly, the Rasch "fit" statistic (total mean square fit) was examined for each item. Over 92% of the items were found to "fit" the model (the fit criterion was set conservatively as a total mean square of less than 1.20). A second type of check was of the slope of the item characteristic curve. The Rasch model assumes that all items have equal discrimination. Items with slope indices (discrimination values) below 0.80 or above 1.20 were defined as being inconsistent with the assumptions of the model. Table II-4 summarizes the number of items outside these bounds.

Table II-4 also summarizes the number of items with low classical discrimination values. These values were the point-biserial indices shown in Table II-3.

Finally, results of a statistical analysis of potential item bias using the Rasch model are summarized. Rasch item difficulty values were computed separately by ethnic group (White, Black and Hispanic) and by sex. These values were then compared and significant differences noted. The assumption underlying this analysis is that the rank order of difficulty should be approximately the same among the various groups. That is, difficult items should be approximately equally difficult for all groups and easy items likewise. This model addresses only differences in relative item difficulty, not absolute difficulty differences.

Statistical analysis of potential item bias is far from a science. However, the techniques used for the HSPT are professionally accepted and frequently used ones. Results of these analyses were not used to revise or remove items from the tests. They were used to highlight statistical anomalies for the test development and minority review committees, whose professional judgments were then used to delete or revise items or to leave them unrevised.

The test development and minority review panels met to examine the Field Test results and to select items to be used on the HSPT. During these reviews, some items were revised or omitted from the tests. Revisions were generally based on the results of the Field Test. Table II-5 summarizes these revisions. Several of the major revisions involved a rewrite of the item stem and options; most Mathematics revisions included a change in the numbers used for solving a problem; two other major changes involved a revision in artwork.

"Minor" revisions included editorial or wording changes in the item stem or an option. It can be seen that revisions were required for fewer than 16% of the items selected. Table II-5 also summarizes the number of items deleted from the tests. Most of these deletions involved omitting excess field-tested items for selected objectives.

Before finalizing the HSPT, one additional set of data was considered. These data involved the speededness of the tests. For practical reasons, nearly all tests have to be administered with time limits. However, achievement tests of the variety represented by the HSPT are not intended to be speeded. That is, although there is a time limit, it should be set so that almost all students have an opportunity to complete the test. In order to insure that this occurs, an analysis of omitted items is made. These data are shown for the Field Test sample in Table II-6. This table summarizes the percent of the sample omitting or failing to reach each test item. This table shows a minor speededness effect in the Writing and Mathematics tests, but a more significant factor in the Reading portion. The speededness affected the Hispanic subsample most significantly, on one Field Test form. This factor was dealt with in the final HSPT by ensuring that the final tests were no longer than earlier years of the HSPT, on which speededness had not been a factor. That these adjustments were sufficient is shown in comparable data in the next section of this manual.

Table II-1

FIELD TEST SAMPLE SIZES BY TEST, FORM, ETHNIC GROUP AND SEX

Test	Form	Total Sample	Sample Size by Ethnic Group			Sample Size by Sex	
			Blacks	Whites	Hispanics	Males	Females
Reading							
	11	350	30	297	15	176	169
	12	340	39	282	9	146	189
	13	342	29	287	14	163	172
	14	344	41	286	9	175	164
	15	348	36	287	11	170	175
	16	339	34	288	8	156	177
Total Number		2063	209	1727	66	986	1046
*Percent			10%	86%	3%	49%	51%
Mathematics							
	21	403	40	344	7	190	211
	22	406	26	365	7	182	219
	23	388	34	341	4	198	189
	24	393	35	343	4	182	206
	25	382	39	322	10	186	194
Total Number		1972	174	1715	32	938	1019
*Percent			9%	89%	2%	48%	52%
Writing							
	31	326	62	241	19	148	164
	32	333	65	237	27	157	165
	33	342	54	260	25	164	165
	34	325	61	240	19	164	146
Total Number		1326	242	978	90	633	640
Percent			18%	75%	7%	50%	50%
Total Sample-Number		5361	625	4420	188	2557	2705
*Percent			12%	84%	4%	49%	51%

*Percent of those who coded ethnic group or sex.

Table II-2

HSPT RAW SCORES FOR FIELD TEST SAMPLES
ACCORDING TO FIELD TEST FORM TAKEN

READING			MATHEMATICS			WRITING		
Field Test Form	HSPT Reading Test Mean	S.D.	Field Test Form	HSPT Math Test Mean	S.D.	Field Test Form	HSPT Writing Test Mean	S.D.
11	77.1	13.5	21	70.0	18.7	31	52.5	8.2
12	77.2	13.3	22	72.7	17.7	32	53.3	8.2
13	75.7	13.9	23	71.2	18.2	33	53.1	8.2
14	75.7	13.5	24	70.8	19.0	34	52.8	8.0
I5	77.3	13.9	25	70.3	19.4			
16	76.3	13.8						

Table II-3

SUMMARY FIELD TEST DATA
BY CONTENT AREA -- ITEM DIFFICULTY AND DISCRIMINATION

Item Difficulty (p) Value	NUMBER OF ITEMS		
	Reading	Writing	Mathematics
.90 - .99	17	25	7
.80 - .89	32	21	20
.70 - .79	22	18	24
.60 - .69	25	11	31
.50 - .59	15	5	21
.40 - .49	11	1	8
.30 - .39	4	3	10
.20 - .29	5	1	3
.10 - .19	1		1
Median p value	.72	.81	.67
Mean p value	.69	.77	.65
Standard Deviation of p values	.19	.16	.17
Number of Items	132	85	125*

Point-Biserial Discrimination Index (r with Total)	NUMBER OF ITEMS		
	Reading	Writing	Mathematics
.60 - .69		4	9
.50 - .59	36	20	31
.40 - .49	48	37	52
.30 - .39	37	18	21
.20 - .29	9	6	10
.10 - .19	1		2
Below .10	1		
Median Point-Biserial	.44	.43	.45

* 4 Mathematics items were field tested in a separate program.

Table II-4

SUMMARY OF FIELD TEST DATA ANALYSIS OF ITEMS
WITH SELECTED STATISTICAL CHARACTERISTICS

DATA QUALITY TESTED*	TEST		
	Reading	Writing	Mathematics
Rasch Model			
Misfit with model	6	7	13
Discrimination - Low	23	18	25
High	31	13	29
Traditional Data			
Low Discrimination	11	6	12
Potential Bias			
Sex - Against Males	20	16	20
Against Females	22	11	17
Ethnic			
Against Hispanics	30	19	+
Against Blacks	28	25	26
Total Number of Items Tested	132	85	125

*Definitions:
Misfit = Rasch total mean square fit value 1.20
Rasch Low Discrimination = Slope of item characteristic curve 0.8
High Discrimination = Slope of item characteristic curve 1.2
(Applied when the item difficulty value was between .25 and .85)

Potential Bias = Rasch difficulty values determined separately by group differ by more than 0.40 logit.

Traditional Low Discrimination = Point-biserial value below .30

+ The Hispanic field test sample was too small to develop stable indices of potential bias.

Table II-5

SUMMARY OF ITEM REVISIONS FOLLOWING
FIELD TESTING--BY TEST AND TYPE OF CHANGE

Content Area	Number of Items in Field Test	Revisions			Items Eliminated*			Total Number of Items in HSPT
		Major	Minor	Total	Inapp.	Extra	Obj.	
Reading-Items	132	12	3	15	1	35	0	96
Passages	17	0	5	5	0	1	0	16
Writing-Mult. Choice	85	0	2	2	0	19	0	66
Mathematics	129	13	12	25	0	18	4	107

*Inappropriate - items eliminated by committees as being an inappropriate measure of the objective.

Extra - Items eliminated because excess items were field tested for a given objective.

Objective - Items from a Mathematics objective that was field tested, but omitted from the HSPT.

Table II-6

PERCENTAGE OF OMITS IN FIELD TEST ITEMS
BY TEST AND ETHNIC GROUP

Percent Omits	NUMBER OF ITEMS											
	READING				WRITING				MATHEMATICS			
	Group				Group				Group			
	Total	White	Hispanic	Black	Total	White	Hispanic	Black	Total	White	Hispanic	Black
Over 10%			2	1								
9-10%			1	1								
7- 8%			2	1								
5- 6%				7			3	1				3
3- 4%	1			24			1	1				14
1- 2%	27	17		3	12	8	0	12	37	35		7
0%	104	115	127	95	73	77	81	71	88	90	+	101
Number of Items	132				85				125*			

* 4 items were tested in a separate program.

+ The Math Hispanic sample was too small to yield stable data.

SECTION III

HSPT Results

The HSPT was administered statewide to all public-school Grade 9 students on April 15-17, 1986. Makeup testing was scheduled for the two-week period immediately following these dates. Table III-1 shows the number of students tested within the District Factor Groups used by the State Department of Education. Table III-2 summarizes the statewide enrollment, the number and percent of students tested by test, the number of students identified by special coding as having special education placement or limited English-speaking ability, and the number of voided or unscorable tests. It can be seen from this table that over 96% of the enrolled, eligible students were actually tested and scored.

Technical data reported in the following tables were developed based on a systematic sample of every tenth student in the statewide data file. This resulted in a sample per test of 8011 students. Table III-3 presents summary data on the comparability of this subsample with the complete statewide population. It is obvious that the subsample provides a very close representation to the population in terms of these measures.

Table III-4 presents a distribution by test cluster of the percent of students receiving each possible raw score; corresponding summary data are also presented. A broad distribution of student performance is obvious in most of the clusters.

A frequency distribution of the item p-values (difficulty values) and item discrimination indices is provided by cluster and test for the Reading, Writing and Mathematics tests in Tables III-5a, III-5b and III-5c, respectively. The top section of each table shows the distribution of item difficulty values.

Two types of item discrimination indices are shown in the bottom portion of the table. The first column of each cluster presents the correlation of the item score with the cluster score--i.e., the relationship of the item with the items within its particular content cluster. The second set of data shows the correlation of the item with the total test score. Both types of indices are point-biserial correlation coefficients. In all correlations, the item score is part of the "total" (cluster or test) to which it is correlated. Partly for this reason, the item-cluster correlations are somewhat higher than the corresponding item-test values.

Raw score frequency distributions for each of the tests are summarized in Table III-6. These data demonstrate the negative skew of the test distributions, especially in Writing and Reading. The distributions evidence both the sufficient "floor" and "ceiling" of the HSPT tests -- that is, almost all students were able to answer correctly a significant proportion of the items, but few students were able to answer all or nearly all items correctly. Test summary data are also presented in this table.

Table III-7 provides summary data and reliability information by cluster. The reliability data reported are Kuder-Richardson Formula 20 estimates, which provide an estimate of the internal consistency of the tests or clusters. As would be expected of long, professionally constructed tests of achievement, the test scores have high reliability coefficients, ranging from .897 in Writing to .961 in Mathematics. Thus, total test scores derived from the HSPT are highly reliable. Reliability is partially a function of test length; therefore, the reliability of the clusters varies rather widely--from .688 for a nine-item cluster to .881 for a 39-item cluster.

For interpreting the reliability of scores of individual students, the standard error of measurement, reported in the last column of this table, is more meaningful than the reliability coefficient. The raw score standard errors of measurement are roughly 3-4 points for the test scores and 1-2 points for the clusters. The interpretation of these indices at the test level is that chances are 2 in 3 that a student's "true" achievement level (were the student tested without errors of any type) would be within 3 to 4 points plus or minus of the obtained score. Thus, if a student obtained a raw score of 40 on the Writing test, chances are 2 in 3 that this "true" score is approximately between 37 and 43. For a 95% degree of confidence, the tabled standard error would be doubled.

All HSPT items were analyzed according to the Rasch model, discussed in Section II of this Technical Report. The State Department of Education uses this mathematical model to equate successive versions of the HSPT. Thus, it is important to determine whether the tests can appropriately be analyzed according to these statistical procedures. Two statistical tests of the appropriateness of the model to the HSPT items were made--fit and discrimination (discussed in Section II). Using the fit test, 254 of 269 items were found to "fit" the model--that is, to have a total mean square value of 1.20 or lower. The misfitting items were not used for test equating purposes. It is important to note that the 1.20 value chosen to define misfitting items is rather conservative. Using this low cutoff increases the number of misfitting items identified. This value was chosen to ensure that all somewhat-atypical items were excluded from the equating process. Approximately 45% of the items had discrimination (slope) values outside the rather narrow band of 0.8 - 1.2. This proportion of items with non-unitary slope indices is consistent with data usually reported for tests such as these. See Table III-8 for a summary of the Rasch data.

The three HSPT tests were each administered in three separately timed parts. Although the test parts were timed, they were essentially "unspeeded" since almost all students reached every item. Table III-9 summarizes these data. The columns especially pertinent in this regard are the last two--"Number Omitted" and "Number Not Reached." "Omitted" items were defined as those prior in sequence to the last marked item for which no response was made. "Not Reached" items were unmarked items that appeared after the last marked item. As is shown in the table, the highest mean number of "omitted" items in any section was 0.13, while the highest mean number of "not reached" items was 0.07. Clearly, almost no students experienced any difficulty completing all parts of the tests.

Table III-10 presents an intercorrelation matrix for the HSPT by test and cluster. The correlations among the three total test scores were as follows:

Reading with Multiple-Choice Writing	= .84
Reading with Total Writing (including the Essay score)	= .72
Reading with Mathematics	= .77
Multiple-Choice Writing with Mathematics	= .76
Total Writing with Mathematics	= .63

These values are consistent with those generally reported for tests of this type for this age level as well as those from earlier years of HSPT testing. Correlations among clusters are partially a function of test length/reliability and are, therefore, difficult to interpret without considering the number of items making up the cluster. Correlations of clusters with the total scores are confounded by the presence of the cluster within the total score. Thus, as would be anticipated because of both of the above factors, cluster-total correlations are largely a function of the number of items in the cluster.

Table III-1

TOTAL POPULATION OF STUDENTS TESTED STATEWIDE
BY DISTRICT FACTOR GROUPING

District Factor Group	NUMBER OF STUDENTS		
	Reading	*Writing	Mathematics
A	14,134	14,003	14,096
B	6,901	6,890	6,901
C	5,465	5,459	5,467
D	6,004	5,999	6,003
E	4,970	4,971	4,969
F	8,097	8,101	8,100
G	9,304	9,307	9,307
H	9,201	9,203	9,203
I	8,445	8,444	8,442
J	5,454	5,450	5,450
V	2,573	2,566	2,573
TOTALS	80,548	80,393	80,511

*Total Writing Scores

Table III-2

SUMMARY OF NUMBER OF STUDENTS TESTED OR EXCLUDED
BY TEST AREA ON THE HSPT

	READING	WRITING			MATHEMATICS
		Total	Essay	Mult-choice	
Enrollment[1]	83,402	83,402	83,402	83,402	83,402
Number Tested[2]	80,773	80,773	80,773	80,773	80,773
Percent of Enrollment Tested	96.8%	96.8%	96.8%	96.8%	96.8%
Special Education Students Tested	4258	4258	4258	4258	4258
Limited English-Speaking Students Tested	1945	1945	1945	1945	1945
Voided Answer Sheet	226		52	357	262
Unscorable Essays			486		
TOTAL NUMBER TESTED	86,968	86,968	86,968	86,968	86,968

[1] Excludes Special Education and limited English-speaking students.

[2] Excludes Special Education and limited English-speaking students and voided answer sheets.

Table III-3

SUMMARY DATA FOR THE TOTAL STATEWIDE POPULATION
AND THE SUBSAMPLE USED FOR SUBSEQUENT STATISTICAL ANALYSES

Test	Subsample	Total Population
READING		
Mean Raw Score	73.14	72.96
Standard Deviation	14.95	15.03
Sample Size	8011	80,548
*WRITING		
Mean Raw Score	53.08	53.01
Standard Deviation	8.86	8.88
Sample Size	8011	80,416
MATHEMATICS		
Mean Raw Score	75.69	75.51
Standard Deviation	20.66	20.44
Sample Size	8011	80,511

*Multiple-Choice Score Only

Table III-4

RAW SCORE FREQUENCY DISTRIBUTIONS AND SUMMARY DATA BY TEST CLUSTER--
PERCENT OF STUDENTS RECEIVING EACH POSSIBLE HSPT RAW SCORE

	READING			WRITING			MATHEMATICS						
Raw Score	LITERAL COMP.	INFER'L COMP.	STUDY SKILLS	SENTENCE STRUCTURE	ORGAN. OF IDEAS	EDIT.	FRACT.	DECIMAL	PERCENT	CONCEPT	MEASURE & GEOM.	PRE-ALGEBRA	PROB. SOLV.
39		3.7											
38		6.7											
37		8.5											
36		8.6											
35		8.5											
34		8.0											
33	1.0	7.1											
32	3.8	5.7											
31	5.1	5.3											
30	7.0	4.3		5.0									
29	8.1	4.0		7.9									3.3
28	8.1	3.7		10.4									5.9
27	7.4	3.4		10.8									6.4
26	7.3	3.1		9.8									6.3
25	6.6	2.7		9.9									5.9
24	6.2	2.3	3.8	9.0	6.6								6.0
23	5.4	2.1	8.0	8.2	12.3								5.3
22	5.6	1.6	10.4	6.8	14.1								5.3
21	4.1	1.8	11.3	5.7	13.1								5.3
20	3.8	1.4	11.2	4.3	12.2						4.1		4.6
19	3.2	1.4	10.0	3.7	9.6						6.5		4.4
18	3.3	1.0	8.9	2.4	7.1						6.5		4.8
17	2.7	1.0	7.1	1.7	5.7						7.2		3.7
16	2.4	.9	6.5	1.3	5.0					9.3	7.1		4.4
15	2.0	.7	5.1	.9	3.2		23.1			15.0	7.3		4.3
14	1.6	.7	4.2	.8	2.8		19.0			15.4	6.5		3.7
13	1.4	.5	4.0	.5	2.0		14.3			12.8	7.3		3.2
12	1.1	.3	2.8	.2	1.5	21.2	9.6			10.2	7.2	17.7	3.5
11	.8	.3	2.2	.1	1.2	20.9	7.9			8.1	6.8	14.6	3.2
10	.5	.3	1.6	.1	1.0	16.7	5.6			6.6	6.6	10.9	2.9
9	.4	.2	.9	.2	.9	12.1	4.5	32.2		5.8	6.2	8.1	2.2
8	.4	.1	.7	.1	.5	9.4	3.6	27.2		5.0	5.5	7.9	2.0
7	.2	.1	.5		.4	7.7	2.9	15.7		4.4	4.7	7.3	1.2
6	.1		.3		.3	5.1	2.4	9.3	38.8	2.9	4.4	7.5	.9
5	.1		.2		.2	3.5	2.2	6.5	17.6	2.0	3.0	7.8	.7
4	.1		.1		.1	1.8	1.9	4.2	13.1	1.4	1.9	6.7	.4
3						1.1	1.4	2.6	12.0	.6	.8	5.5	.1
2						.4	1.1	1.5	10.3	.4	.3	3.8	
1						.1	.5	.8	6.4	.1	.1	1.8	
0						.1	.1	.1	1.7			.4	
Mean	24.12	30.81	18.21	24.27	19.32	9.48	11.85	7.34	4.36	11.91	12.65	8.06	19.52
S.D.	5.60	6.51	3.96	3.98	3.72	2.29	3.32	1.82	1.73	3.22	4.47	3.27	6.36
Mean p Value	.73	.79	.76	.81	.80	.79	.79	.82	.73	.74	.63	.67	.67

Table III-5a

FREQUENCY DISTRIBUTIONS OF ITEM DIFFICULTY VALUES
AND POINT-BISERIAL DISCRIMINATION INDICES BY CLUSTER

HSPT READING

p-Value	Literal Comprehension	Inferential Comprehension	Study Skills	Total
.950-.999		1	2	3
.900-.949	3	4	1	8
.850-.899	5	9	5	19
.800-.849	6	8	5	19
.750-.799	5	5	1	11
.700-.749	3	6	1	10
.650-.699	3	3	3	9
.600-.649	3	1	1	5
.550-.599	1	1	2	4
.500-.549	0		1	1
.450-.499	0		1	1
.400-.449	2		1	3
.350-.399		1		1
.300-.349	1			1
.250-.299	1			1
Mean p-Value	.73	.79	.76	.76
Median p-Value	.79	.81	.81	.79

Point-Biserial Correlation	Cluster	Total	Cluster	Total	Cluster	Total	Cluster	Total
.70 - .79								
.60 - .69								
.50 - .59	6	4	6		5	1	17	5
.40 - .49	13	13	22	25	10	12	45	50
.30 - .39	10	11	9	9	6	4	25	24
.20 - .29	4	5	2	5	2	6	8	16
.10 - .19					1	0	1	0
.01 - .09						1	0	1
Median	.42	.40	.44	.42	.43	.41	.44	.42
Total Number of Items	33		39		24		96	

Table III-5b

FREQUENCY DISTRIBUTIONS OF ITEM DIFFICULTY VALUES
AND POINT-BISERIAL DISCRIMINATION INDICES BY WRITING

HSPT WRITING

p-Value	Sentence Structure	Organization of Ideas	Editing	Total
.950-.999	6	1		7
.900-.949	7	5	1	13
.850-.899	2	7	3	12
.800-.849	3	2	4	9
.750-.799	4	0	1	5
.700-.749	0	4	1	5
.650-.699	3	2	1	6
.600-.649	1	2	0	3
.550-.599	2	1	1	4
.500-.549	0			0
.450-.499	1			1
.400-.449	0			0
.350-.399	0			0
.300-.349	1			1
Mean p-Value	.81	.80	.79	.80
Median p-Value	.86	.85	.81	.84

Point-Biserial Correlation	Cluster	Total	Cluster	Total	Cluster	Total	Cluster	Total
.50 - .59	3	1	2		6	1	11	2
.40 - .49	7	8	11	10	4	5	22	23
.30 - .39	14	13	11	12	2	4	27	29
.20 - .29	6	8		2		2	6	12
Median	.36	.32	.41	.35	.50	.41	.40	.35
Total Number of Items	30		24		12		66	

Table III-5c

FREQUENCY DISTRIBUTIONS OF ITEM DIFFICULTY VALUES
AND POINT-BISERIAL DISCRIMINATION INDICES BY CLUSTER

HSPT MATHEMATICS

p-Value	Computation: Fractions	Computation: Decimals	Computation: Percents	Concepts	Measure. & Geometry	Pre-Algebra	Problem Solving	TOTAL
.950-.999		1		1				2
.900-.949	1	0		1				2
.850-.899	4	3		2	2		3	14
.800-.849	2	1		1	1	1	1	7
.750-.799	3	1	2	4	3	2	4	19
.700-.749	3	2	2	2	0	2	3	14
.650-.699	2	1	2	2	2	1	2	12
.600-.649				1	5	2	9	17
.550-.599				0	2	3	3	8
.500-.549				1	0	1	2	4
.450-.499				1	1		2	4
.400-.449					2			2
.350-.399					2			2
Mean p-Value	.79	.82	.73	.65	.63	.67	.67	.71
Median p-Value	.78	.84	.73	.76	.64	.68	.64	.72

Point-Biserial Correlation	Fractions Clus.	Fractions T	Decimals Clus.	Decimals T	Percents Clus.	Percents T	Concepts Clus.	Concepts T	Measure. & Geometry Clus.	Measure. & Geometry T	Pre-Algebra Clus.	Pre-Algebra T	Problem Solving Clus.	Problem Solving T	TOTAL Clus.	TOTAL T
.70 - .79					1										1	
.60 - .69	2		2		4		3	1	1		6				18	1
.50 - .59	11		5	1	1	3	4	3	10	4	5	6	13	2	49	19
.40 - .49	2	13	1	3		2	4	5	5	10	1	6	13	23	26	62
.30 - .39		2		4		1	5	3	3	4			3	1	11	15
.20 - .29			1	0				4	1	1				3	2	8
.10 - .19				1						1						2
Median	.56	.45	.56	.38	.66	.50	.49	.43	.53	.46	.60	.50	.49	.46	.52	.46
Total Number of Items	15		9		6		16		20		12		29		107	

Table III-6

RAW SCORE FREQUENCY DISTRIBUTIONS
AND SUMMARY DATA BY TEST
PERCENT OF STUDENTS RECEIVING EACH POSSIBLE HSPT RAW SCORE

Raw Score	TEST		
	Reading	Writing	Mathematics
106-107			1.1
103-105			4.5
100-102			6.0
97-99			6.6
94-96	1.4		6.2
91-93	6.0		5.9
88-90	8.6		5.7
85-87	10.2		5.4
82-84	9.4		5.1
79-81	8.8		4.5
76-78	8.1		4.5
73-75	7.2		4.6
70-72	5.9		4.0
67-69	5.3		4.3
64-66	4.9	7.1	3.7
61-63	4.2	14.2	3.5
58-60	3.4	16.1	3.1
55-57	3.3	14.8	3.0
52-54	2.6	12.3	3.0
49-51	2.5	9.5	2.5
46-48	1.9	7.7	2.6
43-45	1.6	5.9	2.2
40-42	1.4	4.6	1.9
37-39	.9	2.6	1.4
34-36	.7	2.1	1.6
31-33	.5	1.3	1.3
28-30	.4	.7	.8
25-27	.2	.5	.5
22-24	.2	.5	.3
19-21	.1	.1	.2
16-18	.0	.2	.1
13-15	.0	.0	.0
10-12	.0	.0	.0
7-9			.0
4-6			.0
0-3			.0
Mean	73.14	53.08	75.69
S.D.	14.95	8.86	20.66
Median	76.48	54.94	79.20
No. Items	96	66	107

Table III-7

HSPT SUMMARY DATA AND RELIABILITY ESTIMATES BY CLUSTER AND TEST

Test/Cluster	Number of Items	RAW SCORE Mean	RAW SCORE S.D.	K-R 20 Reliability	Standard Error of Measurement
READING	96	73.1	14.9	.941	3.9
Literal Comprehension	33	24.1	5.6	.848	2.4
Inferential Comprehension	39	30.8	6.5	.881	2.4
Study Skills	24	18.2	4.0	.787	2.0
WRITING	66	53.1	8.9	.897	3.1
Sentence Structure	30	24.3	4.0	.783	2.1
Organization of Ideas	24	19.3	3.7	.783	1.8
Editing	12	9.5	2.3	.702	1.3
MATHEMATICS	107	75.7	20.7	.961	4.3
Fractions	15	11.8	3.3	.839	1.4
Decimals	9	7.3	1.8	.688	1.1
Percents	6	4.4	1.7	.726	0.9
Concepts	16	11.9	3.2	.779	1.6
Measurement & Geometry	20	12.6	4.5	.831	2.0
Pre-Algebra	12	8.1	3.3	.833	1.4
Problem Solving	29	19.5	6.4	.881	2.3

*An estimate of the reliability of the Total Writing score is provided on page 39.

Table III-8

SUMMARY OF THE CONSISTENCY OF HSPT ITEMS
WITH THE ASSUMPTIONS OF THE RASCH MODEL*

DATA CHARACTERISTIC	NUMBER OF ITEMS BY TEST Reading	Writing	Mathematics
Misfit with model	4	2	9
Discrimination (Slope)			
Low	21	13	19
High	31	18	20
Total Number of Items in HSPT	96	66	107

*See text for an explanation of the characteristics.

Table III-9

SUMMARY DATA FOR NUMBER OF ITEMS RIGHT, WRONG, OMITTED OR NOT REACHED* BY PART FOR THE HSPT

Test	No. Part	No. Items	Number Right Mdn	Number Right Mode	Number Right Mean	Number Right S.D.	Number Wrong Mdn	Number Wrong Mode	Number Wrong Mean	Number Wrong S.D.	Number Omitted Mdn	Number Omitted Mode	Number Omitted Mean	Number Omitted S.D.	Number Not Reached Mdn	Number Not Reached Mode	Number Not Reached Mean	Number Not Reached S.D.
READING	1	32	25	27	23.8	5.0	7	5	8.1	4.9	0	0	0.04	0.34	0	0	0.05	0.83
	2	32	26	29	24.5	5.3	6	3	7.4	5.2	0	0	0.05	0.44	0	0	0.07	0.86
	3	32	26	30	24.8	5.7	6	2	7.1	5.6	0	0	0.06	0.42	0	0	0.03	0.71
WRITING**	2	36	30	33	29.1	5.0	6	3	6.9	5.0	0	0	0.03	0.24	0	0	0.02	0.40
	3	30	25	27	24.0	4.4	5	3	6.0	4.3	0	0	0.04	0.31	0	0	0.01	0.51
MATHEMATICS	1	35	26	31	25.2	6.6	9	4	9.7	6.4	0	0	0.10	0.71	0	0	0.05	0.75
	2	35	26	33	24.9	7.4	8	2	10.0	7.3	0	0	0.10	0.53	0	0	0.05	0.99
	3	37	27	32	25.7	7.5	10	4	11.2	7.4	0	0	0.13	0.61	0	0	0.06	1.12

* See text for a definition of "Omitted" and "Not Reached" items.

** Part 1 was the Essay.

Table III-10

INTERCORRELATION MATRIX BY CLUSTER AND TEST

Test/Cluster	Reading			Mathematics								Writing					
	R2	R3	RT	M1	M2	M3	M4	M5	M6	M7	MT	W1	W2	W3	W4	WMC	WT
READING																	
R1 Literal Comp.	.86	.75	.95	.50	.49	.53	.63	.66	.62	.69	.72	.72	.74	.60	.45	.79	.67
R2 Inferential Comp.		.74	.95	.50	.48	.51	.61	.63	.60	.68	.70	.72	.75	.59	.44	.79	.67
R3 Study Skills			.87	.54	.54	.54	.66	.67	.64	.71	.75	.67	.68	.63	.42	.75	.64
RT Total				.55	.53	.56	.67	.70	.66	.74	.77	.76	.78	.65	.47	.84	.72
MATHEMATICS																	
M1 Comp. - Fractions					.66	.55	.70	.60	.61	.64	.80	.51	.50	.52	.31	.57	.48
M2 Comp. - Decimals						.52	.67	.58	.58	.60	.75	.48	.49	.50	.29	.55	.46
M3 Comp. - Percents							.64	.62	.57	.64	.74	.50	.48	.46	.30	.54	.46
M4 Number Concepts								.73	.70	.75	.88	.60	.59	.58	.35	.67	.55
M5 Meas & Geometry									.71	.78	.88	.61	.59	.56	.37	.67	.56
M6 Pre-Algebra										.72	.84	.60	.58	.59	.37	.66	.56
M7 Problem Solving											.91	.64	.63	.58	.38	.70	.59
MT Total												.68	.67	.65	.42	.76	.63
WRITING																	
W1 Sentence Structure													.72	.64	.44	.92	.73
W2 Organization														.62	.42	.90	.71
W3 Editing															.39	.81	.65
W4 Essay																.47	.91
WMC MC Total																	.79
WT Total																	

SECTION IV

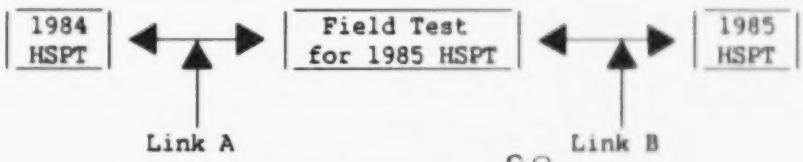
Equating

Each year, totally new test items are developed for the HSPT. The new items, while assessing a uniform set of skills each year, can vary in difficulty when compared with the items assessing the same skill in previous years. Because the item difficulties can vary, the difficulty of the various clusters and the total test can also vary from year to year.

If each new form of the test could be slightly easier or more difficult than earlier versions, it would not be possible to accurately compare results or assess performance across years. Further, and more importantly, students would have either a greater or lesser likelihood of passing each test depending on the difficulty of the version taken. To allow comparison of results from year to year, and to make certain that students have an equal chance of passing the tests regardless of which year's test they take, a statistical procedure known as equating is used. Equated scores can be compared from year to year because they represent equal levels of achievement based upon the score distributions obtained in the first year of HSPT administration. The equated scores are reported using a score scale developed for the program. The HSPT scale range is from zero to 100 for the Reading, Mathematics and Writing (multiple-choice section). The total Writing test scores are reported on a scale of 40 to 100.

Test Equating. Several statistical procedures for test equating have found general acceptance over the years. On the Minimum Basic Skills (MBS) test, for example, a procedure known as linear equating was used. In recent years, many test developing agencies have begun to use one of various procedures known as item response theory models for equating purposes. There are several such procedures, the most widely used of which is the one-parameter or Rasch model. This model has found wide acceptance primarily because of its technical strengths, but also because it is a reasonably straightforward procedure to apply. It was chosen by the N.J. Department of Education, with the advice of various consultants including the Technical Advisory Committee, composed of nationally recognized experts in statistics and measurement. The following outlines the equating procedures used to equate the 1986 tests to the 1985 HSPT and back to the "Base" 1984 tests.

The HSPT was administered for the first time in March, 1984. The 1985 HSPT was constructed from test items field tested in Fall, 1984, as explained in Section II of this report. The equating "link" between the 1984 HSPT (base) test and the 1985 HSPT was made via the Fall, 1984 Field Test. This was accomplished by imbedding a subset of 1984 HSPT items by content area in the Field Test booklets. Thus, the students in the Field Test samples took portions of both the 1984 and 1985 HSPTs. The test equating was conducted using these sets of overlapping items in two stages as shown below:



Link A equated the 1984 HSPT to the 1985 Field Test. A subset of "anchor" or "common" items between the two tests was used to accomplish this. Using procedures outlined in the 1984-85 HSPT Technical Report, Rasch logit constants were calculated to link the two sets of tests.

Link B equated the 1985 Field Test to the 1985 HSPT using procedures also outlined in last year's Technical Report. The result of these operations was an equivalence relationship between the 1985 HSPT and the 1984 HSPT.

During the 1985-86 year, two comparable links were required--first to equate the 1986 Field Test back to the 1985 HSPT (Link C), then to equate the 1986 Field Test to the 1986 HSPT (Link D). The procedures followed are outlined below.

Link C involved the following steps, taken separately for the Reading, Writing, and Mathematics Tests:

1. Rasch model item difficulties were obtained for the 1985 HSPT items using a stratified sample of approximately 6300 students representative of the statewide Grade 9 population. Mean Rasch difficulty values were calculated by test.

2. Rasch analyses were conducted separately for each of the 15 Field Test forms (6 in Reading, 5 in Mathematics, and 4 in Reading). In these analyses, the complete 1985 HSPT content area test plus the Field Test items of the appropriate form were considered as a single, combined test. All items in step 1 or 2 above that had Rasch mean square fit values in excess of 1.20 were eliminated from the following steps.

3. Mean Rasch difficulty values were calculated by form for all items (HSPT plus Field Test), excluding those misfitting items referred to above. The derived mean Rasch logit values for these items were:

READING		MATHEMATICS		WRITING	
Form 11	-0.187	Form 21	-0.028	Form 31	+0.186
Form 12	-0.049	Form 22	-0.020	Form 32	-0.216
Form 13	-0.117	Form 23	-0.009	Form 33	+0.011
Form 14	-0.174	Form 24	-0.065	Form 34	-0.142
Form 15	+0.073	Form 25	-0.133		
Form 16	-0.132				

4. Field Test forms were equated to each other and, concurrently, to the 1985 (and 1984) HSPT using constants derived by subtracting the above values from the constants for the corresponding test as calculated in step 1 above.

Link D was conducted to equate the 1986 Field Tests to the 1986 HSPT. The following steps were required to determine these linking constants.

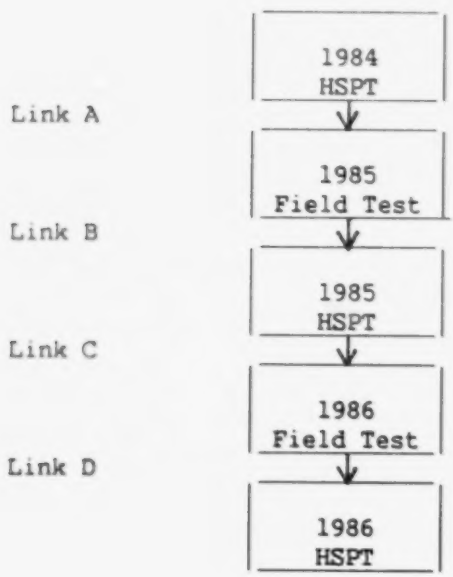
1. Mean Rasch difficulty values were calculated by tryout form for 1986 Field Test items that were not changed between field testing and use in the 1986 HSPT. Also excluded from these calculations were any Field Test items with Rasch mean square fit values of 1.20 or higher. A total of 77 Reading (excluding 12 revised items and 7 misfits), 62 Mathematics (excluding 13 revised and 12 misfitting items), and 63 Writing (excluding 2 revised and 1 misfitting items) items remained after these eliminations for use in equating.

2. Field Test forms-equating constants calculated for Link C were added to the item difficulties for the appropriate form. This resulted in all HSPT items being placed on the same Rasch difficulty scale. A forms-combined mean Rasch item difficulty value was calculated for each content area test.

3. Using a stratified sample of 5887 students selected from the total statewide HSPT distribution, Rasch item difficulty values were computed for each 1986 HSPT item. Mean item difficulty was calculated for the 77 Reading, 82 Mathematics, and 63 Writing items that were neither changed after field testing nor showed a Rasch misfit in either the Field Test or HSPT.

4. By test, the mean HSPT item difficulty value determined in step 3 was subtracted from the Field Test mean difficulty in step 2. The resulting constants were added to each of the 1986 HSPT items in the respective tests. This calibrated the 1986 HSPT to the 1986 Field Test difficulty scale, which (through Link C) had already been equated to the 1984 HSPT.

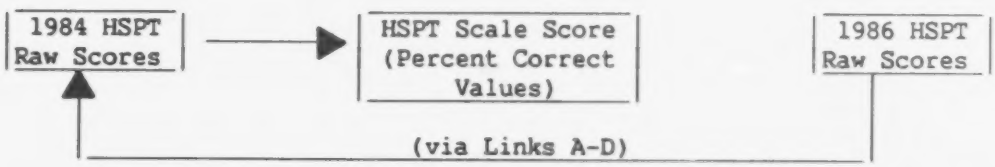
The total equating process to date is illustrated below.



Score Scale

A score scale was developed for the HSPT for purposes of interpretation. The scale essentially represents the percent of items answered correctly on the 1984 HSPT. Each 1984 HSPT raw score was converted to a percent-correct value. For example, a raw score of 33 on the 66-item Writing test converted to a HSPT scale score of 50.

After the 1986 HSPT was equated back to the 1984 tests as above, the 1986 raw scores were expressed in terms of equated 1984 raw scores, and then converted to the HSPT scale score. These procedures are shown below:



In the process of equating the 1986 HSPT raw scores to the 1984 raw scores, the HSPT scale scores were also equated to the 1986 raw scores. All reporting of HSPT results in terms of these equated scale values so that score changes across years can appropriately be assessed. That is, simple raw score comparisons across years cannot be made appropriately because the tests will likely vary slightly in difficulty from year to year. After the raw score equating has been conducted, comparisons of raw score changes across pairs of years could appropriately be made; however, raw score comparisons typically have little interpretive meaning. Percent-correct values, as the HSPT scale values are, have a degree of interpretive meaningfulness that makes them more useful as an across-year score metric.

The HSPT scales developed for the total tests were similarly developed for each test cluster (3 in Reading and Writing, 7 in Mathematics). Thus, both HSPT test scores and cluster scores can be compared from year to year in a similar fashion. In each year of the HSPT program, the results are equated back to the original 1984 base year to serve as a constant benchmark for assessing score change. Table IV-1 presents the conversion table from 1986 raw scores to the HSPT scores.

Comparative HSPT results for 1985 and 1986 are presented in terms of HSPT scores in the document State Report: 1985-86, Analysis and Interpretation of Statewide Performance. Interested readers should request a copy of this document from the Bureau of Cognitive Skills.

TABLE IV-1

1986 RAW SCORE TO HSPT SCORE CONVERSION TABLES

READING						MATHEMATICS						WRITING			
Raw Score	HSPT Score	Raw Score	HSPT Score	Raw Score	HSPT Score	Raw Score	HPST Score	Raw Score	HSPT Score	Raw Score	HSPT Score	Raw Score	HSPT Score	Raw Score	HSPT Score
0	0	36	51	72	85	0	0	36	37	72	68	0	0	36	58
1	1	37	53	73	86	1	1	37	38	73	70	1	2	37	59
2	3	38	54	74	87	2	2	38	39	74	70	2	3	38	61
3	5	39	55	75	88	3	3	39	40	75	71	3	5	39	62
4	7	40	56	76	89	4	4	40	41	76	72	4	6	40	64
5	8	41	57	77	89	5	5	41	42	77	73	5	9	41	65
6	9	42	59	78	90	6	7	42	43	78	74	6	11	42	67
7	11	43	60	79	91	7	8	43	43	79	75	7	12	43	68
8	13	44	61	80	92	8	9	44	44	80	76	8	14	44	70
9	15	45	61	81	92	9	10	45	44	81	76	9	15	45	71
10	16	46	62	82	92	10	11	46	45	82	76	10	18	46	73
11	17	47	63	83	93	11	12	47	46	83	77	11	18	47	74
12	20	48	64	84	93	12	13	48	47	84	79	12	20	48	76
13	21	49	66	85	94	13	15	49	48	85	80	13	21	49	76
14	22	50	67	86	95	14	16	50	49	86	81	14	23	50	77
15	24	51	68	87	95	15	17	51	51	87	82	15	24	51	79
16	25	52	69	88	96	16	18	52	52	88	83	16	26	52	80
17	26	53	70	89	97	17	19	53	53	89	83	17	29	53	82
18	28	54	70	90	97	18	19	54	53	90	84	18	30	54	83
19	30	55	71	91	98	19	21	55	54	91	85	19	32	55	86
20	31	56	72	92	98	20	22	56	55	92	86	20	33	56	86
21	32	57	74	93	99	21	23	57	56	93	87	21	35	57	88
22	33	58	74	94	99	22	24	58	56	94	88	22	36	58	89
23	36	59	75	95	99	23	25	59	57	95	89	23	38	59	91
24	37	60	76	96	100	24	26	60	58	96	89	24	39	60	92
25	38	61	77			25	27	61	59	97	90	25	41	61	92
26	39	62	77			26	28	62	60	98	91	26	42	62	94
27	40	63	78			27	28	63	60	99	92	27	44	63	95
28	41	64	79			28	29	64	61	100	94	28	45	64	97
29	44	65	80			29	30	65	62	101	95	29	47	65	98
30	45	66	80			30	31	66	63	102	95	30	48	66	100
31	46	67	82			31	32	67	65	103	96	31	50		
32	47	68	83			32	33	68	66	104	97	32	52		
33	48	69	83			33	34	69	67	105	98	33	53		
34	49	70	84			34	35	70	67	106	98	34	55		
35	50	71	85			35	36	71	68	107	100	35	56		

SECTION V

Essay and Total Writing Test

The HSPT Writing test is comprised of two parts -- a multiple-choice section and an essay section. From the Writing test, students receive a score for the multiple-choice section, a score for the essay section, and a combined score in which the essay score is weighted at sixty percent of the total and the multiple-choice is weighted at forty percent of the total. The preceding sections of this manual detail the development of the multiple-choice part of the HSPT Writing test. The following information describes the development and scoring of the essay writing section.

Specific information on the rationale for the New Jersey State Department of Education's development of the Registered Holistic scoring method, application of the scoring method criteria and details of the entire process are discussed in the document entitled: The Registered Holistic Scoring Method for Scoring Student Essays (NJSDE, 1985; 1986).

In November, 1984 and 1985, small-scale scoring sessions were held in order to select from a number of field-tested topics those topics that could be "banked" or used as prompts in future years for statewide writing assessment. In order to determine the appropriateness of each topic as an essay prompt, the Writing Test Development/Test Interpretation Committee was convened to read, score and evaluate papers from each topic. The following elements were considered as selection criteria to evaluate the appropriateness of a topic.

1. The student understands the topic and produces enough writing to allow for responsible appraisal.

2. The topic allows the student to respond with specific examples and supporting details.

3. The topic elicits a range and variety of responses.

4. The topic provides for a statistically reasonable distribution of scores on a holistic scoring scale of 1 to 6.

5. The topic does not elicit highly emotional responses from either writer or reader.

6. The topic does not appear to favor any single group defined by gender or cultural background.

7. The topic allows the student's writing to exhibit quality, level of cognition, appropriateness and enthusiasm of response.

8. The topic allows for a reasonable degree of scorer agreement.

9. The topic realistically sustains reader interest over a period of time.

During this topic selection process, a pair of readers independently assigned scores from 1 (low) to 6 (high) for each essay. These scores were summed to obtain the total essay score (on a scale of 2 to 12). The range and distribution of essay scores and point differentials for each of the essay forms were presented to the committee to facilitate the group's discussion and final selection of essay topics. Based on the committee's reading and discussion of the forms field tested, the committee agreed to bank several of the essay topics. In 1986, based on the committee's past decision, Field Test Essay Form 1 was the essay topic used in April for the statewide essay prompt.

The selected topic follows:

> If you could go back in time, what day or event of your life would you like to live over? Write an essay that tells about that day or event and what made it special.

The essay topic was administered as Part 1 of the writing test on April 17, 1986. As shown in Table III-2, 86,968 students were tested. Although essays from special education and limited English-speaking students were scored, these scores were not aggregated. In addition, folders which had to be voided were also excluded resulting in 80,773 students tested.

During the essay-scoring process, 486 papers of the 80,773 folders were determined to be unscorable. Of these unscorable papers, 424 papers showed no response, 50 were judged off topic, eight were fragmented and four were not English.

The essay section of the writing test was hand-scored in Durham, North Carolina by 70 qualified educators trained to use the Registered Holistic scoring procedure (NJSDE, 1985). The following is an outline of the criteria used to select readers.

I. Evidence of experience as:

   A. English Department Chairperson or
   B. Teacher of English, Communication, and/or Composition or
   C. Teacher of other disciplines

II. Certification: Bachelor of Science/Arts (minimum)

III. Currently working in the educational community

IV. Participate in two oral interviews

V. Submission of a literate position statement written as part of the application in response to a question about the applicant's experience and motivation

VI. Submission of an essay on a given topic

VII. Experience in large-scale assessment scoring procedures (scoring consistency based on previous job performance)

VIII. Following training, the ability to meet the scoring criteria established as "perfect agreement" for three out of five qualifying sets of essays. (Each qualifying set contains 15 to 20 prescored student essays.)

The Registered Holistic scoring sessions for the HSPT occurred on the following dates: May 13 - June 2, 1986.

The Registered Holistic scoring method is a different way of assessing students' writing. While other scoring methods (holistic, analytic, primary trait) vary with the nature of the writing to be scored (narration, description, etc.) and the local instructional objectives to be measured, Registered Holistic is based on defined features of writing quality and levels of writing proficiency. The Registered Holistic scoring method enables raters to consistently and accurately judge student writing samples when generated from one year to the next in a test setting. Raters trained in using this scoring method judge the sample in terms of how well the writer has integrated four features of written language and produced a writing sample that communicates an intended message to a given audience. Proficiency and success are based upon student command or level of facility with four salient features of written expression: organization/content, usage, sentence construction, and mechanics.

In Registered Holistic scoring, the predetermined set of features (organization/content, usage, sentence construction and mechanics) are registered or anchored by statement(s) that create the scoring guide criteria. The point values or score points describe the proficiency level or how well the student handles the four features. New Jersey uses a six-point score scale: 1 (lowest) to 6 (highest). Each score point on the six-point scale has the same four features but with inherent differences in their descriptions relative to the level of proficiency. The criteria with descriptive features indicate at each score point the student-writer's command of written language.

The Registered Holistic Scoring Method: An Operational Definition

The Registered Holistic scoring method judges students' command of written language. The features of the scoring guide criteria are organization/ content, usage, sentence construction, and mechanics. Criteria for this method are deliberately limited to what are considered the important aspects of writing at the state and local level. These features were considered by experts as most important to a written response that communicates an intended message to an intended audience (Breland & Jones, 1984; Freedman, 1979). The features were distilled from analyses of composition features referenced by expert readers and three years of summary reports conducted by the New Jersey State Department of Education (Bloom, 1985; Bloom & Rabinowitz, 1985). Furthermore, these features can be taught and are considered worthy of instruction. Some programs assess writing using scoring guides with two criteria that are domain specific for text-level features. For example, these scoring guides would direct raters to look for: a) the parts of a business letter; b) focus and organization; and/or c) 10 separate criteria for sentence-level conventions--implying that mechanics are more important than

the focus of content and the organization of related ideas. In contrast, the criteria of the Registered Holistic scoring method are based or anchored by four features that define increasing levels of writing skill development while remaining independent of the age of the writers as well as the domain or mode they choose to use to communicate. This independence of criteria among writer, mode, and prompt is considered a critical element that needs to be addressed within large scale assessment programs that deny diplomas based upon students' ability to command written language in response to a given prompt. Thus, using the Registered Holistic method, the first focus of raters is to the feature of "organization/content," with a readjustment of that score according to the severity of errors or error patterns in usage, sentence construction, and/or mechanics.

The significant difference between this method and most other scoring methods is the criteria's focus and flexibility. The Registered Holistic method's focus is predominantly on organization/content. The method is flexible in that it can be used with any topic prompt and any mode of writing.

The Registered Holistic scoring method:

- uses score point criteria representing increasing levels of writing skill development;
- uses consistent language across all score points;
- defines each feature at each score point, emphasizing content and organization; and
- provides accuracy and consistency through stringent controls on the training and qualifying of raters.

The scoring method does not rely on raters to infer or overlay the criteria/feature from examples or anchor papers. Rather, the method relies on the consistent and defined criteria at each score point.

To establish the scoring process and scoring guide for the Registered Holistic scoring method, a random sampling of the papers was reviewed and the qualifying and training (Q/T) papers were identified. The "Q/T" papers are considered to be exemplary of the four predetermined and defined constructs or features at each of the six score points of the Registered Holistic scoring method. The qualifying and training paper selection session was held in Durham, North Carolina (subcontractor Measurement Incorporated (MI)) on the following dates: April 29 - May 2, 1986. The five participants in this selection process included 4 consultants (3 MI Chief Readers and 1 NJ consultant Chief Reader) and New Jersey's Writing Project Director. Eighty papers were chosen for training the readers. Copies of papers selected for the qualifying/training sets are on file with the department. The training of the 70 readers and the qualification of those readers occurred on the following dates: May 5 - 12, 1986. The actual scoring of the student essays was on May 13 - June 2, 1986.

The consistency between readers and among scores by individual readers is an important aspect of Registered Holistic scoring. This level of reader agreement and score consistency is maintained by the chief reader and team leaders for the duration of the Registered Holistic scoring session. The Chief Reader conducts calibration sessions with all readers using "Q/T" papers three to four times every day. These sessions assure that the reading and scoring of individual essays are based on the individual student's writing ability in relationship to the features defined at each score point of the Registered Holistic scoring guide.

In addition, scoring sets are re-scored, randomly spot-checked by team leaders, and cross-scored by third and fourth team leaders to guard against cross-team drift. This procedure contributes to roomwide consistency.

As indicated above, each essay receives two readings and two separate scores. The second reader has no knowledge of the first reader's score. A six-point scale is used for this evaluation process. A score of 1 represents an "inadequate command" of written language and a score of 6 represents a "strong command". Therefore, a student's essay could receive a score from 2 to 12, the sum of two readers' ratings. Prior to the scoring session, it was decided that when the two scores assigned a given writing sample were discrepant by more than one score point, the essay would be read a third time. In cases of three readings, the HSPT essay score was determined by summing the scores for the three readings and multiplying by two-thirds.

In 1986, the essay components of the examination was taken by 86,968 students.

TABLE V-1

DISTRIBUTION OF HSPT ESSAY SCORES

Percentage of Students	Distribution of Scores											
	0	2	3	4	5	6	7	8	9	10	11	12
	0.6	0.1	0.0	1.3	1.7	18.5	10.7	36.3	14.1	12.8	2.8	1.1

The total writing test score was obtained by combining the multiple-choice score with the essay score, giving the essay a weight of 60 percent of the total. The total test score was reported on a scale of 40 to 100 with a mean of 80 and a standard deviation of 10. The passing score on the total writing test score (multiple choice and essay) was 77.

Using the .612 correlation as a measure of inter-rater agreement, the reliability of the essay test can be computed using the Spearman-Brown Prophecy Formula, since the essay score is computed as the sum of two ratings:

$$r = \frac{2(.612)}{1 + .612} = .759$$

Thus, the reliability of the essay test was .76.

The method used to derive the total score is one which is typically used when scores of different ranges must be combined. The range for the essay test is 2-12; the range for multiple-choice test is percents 0-100. The method uses z-score statistics as follows:

1. The statewide mean and standard deviation of the essay and multiple-choice scores were determined.

	Mean	Standard Deviation
Essay	7.5	2.2
Multiple-choice	76.6	16.1

2. Compute the z-scores for the essay and multiple-choice tests:

$$z\ \text{Essay} = \frac{(\text{Essay Scores} - 7.5)}{2.2}$$

$$z\ \text{Multiple-Choice} = \frac{(\text{Multiple-Choice Score} - 76.6)}{16.1}$$

3. Combine the z scores, using the weighting of 60 percent for the essay score and 40 percent for the multiple-choice score.

$$Xi = 0.6\ (z\ \text{Essay}) + 0.4\ (z\ \text{Multiple-Choice})$$

This is a weighted sum of each student's essay and multiple-choice scores after each score has been converted to a z-score. The weighted sums, Xi, are distributed such that $\bar{X} = 0$ and $S_x = .93$.

4. The weighted sums derived above are then transformed into the total test reporting scale of 40-100 with a mean of 80 and a standard deviation of 10 by the following equation:

$$\text{Total Score} = 80 + 10\ \frac{Xi - \bar{X}}{S_x}$$

$$= 80 + 10\ \frac{Xi}{.93}$$

where .93 is the standard deviation of the scores derived in Step 3.

5. The scoring formula could yield scores greater than 100.0. In these cases, these scores were reported as 100.

The reliability of the Total Writing scale was estimated using the following formula:

$$r_{cc} = \frac{.4^2 (r_{mc,mc}) (s^2_{x_{mc}}) + .6^2 (r_{w,w,}) (s^2_{x_w}) + 2(.4)(.6)(r_{x_{mc}x_w})(s_{x_{mc}})(s_{x_w})}{.4^2 (s^2_{x_{mc}}) + .6^2 (s^2_{x_w}) + 2 (.4)(.6)(r_{x_{mc}x_w})(s_{x_{mc}})(s_{x_w})}$$

Where $r_{c,c}$, $r_{mc,mc}$, $r_{w,w,}$ are the reliabilities of the Total Writing (composite) test (c), the multiple-choice section (mc), and the essay (w), $s^2_{x_c}$, $s^2_{x_{mc}}$, and $s^2_{x_w}$ are the variance of the observed scores on the three measures, and $r_{x_{mc}x_w}$ is the correlation between the multiple-choice and essay portions.

In calculating this reliability of the composite Writing score, the reliability estimate used for the essay was .759, the inter-rater reliability of the essay discussed on page 38.

Using the above formula, the reliability of the Total Writing scale is estimated as .862. The standard error of measurement expressed in scaled score terms is 3.72.

Table V-2 shows the frequency distribution for the total writing test scores.

Table V-2

1986 DISTRIBUTION OF HSPT TOTAL WRITING TEST SCORE

Passing Score = 77 Statewide Mean = 82.2

	Total Test Scores				
Percentage of Students	40-52	53-64	65-76	77-88	89-100
	0.3	2.5	20.7	54.7	21.8